Learning Localized Perceptual Similarity Metrics for Interactive Categorization

Catherine Wah *
Google Inc.
google.com

Subhransu Maji
UMass Amherst
cs.umass.edu

Serge Belongie
Cornell Tech
vision.cornell.edu

Abstract

Current similarity-based approaches to interactive fine-grained categorization rely on learning metrics from holistic perceptual measurements of similarity between objects or images. However, making a single judgment of similarity at the object level can be a difficult or overwhelming task for the human user to perform. Secondly, a single general metric of similarity may not be able to adequately capture the minute differences that discriminate fine-grained categories. In this work, we propose a novel approach to interactive categorization that leverages multiple perceptual similarity metrics learned from localized and roughly aligned regions across images, reporting state-of-the-art results and outperforming methods that use a single nonlocalized similarity metric.

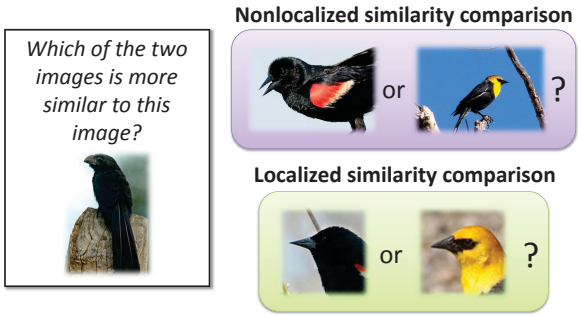


Figure 1. In this work, we use perceptual similarity metrics learned from localized comparisons to perform interactive categorization. By directing the user's attention to localized and roughly aligned regions, we aim to reduce both overall human effort required for categorization as well as improve performance over using nonlocalized comparisons and metrics.

1. Introduction

Fine-grained visual categorization (FGVC) is an area of computer vision that has experienced an increased amount of attention in recent years across various visual domains [39, 9, 27, 52]. The goal is to distinguish between *fine-grained categories* or subcategories (*e.g.*, a Cardinal vs. a Lazuli Bunting) that belong to the same *basic-level category* (*e.g.*, Bird).

Some work has focused on interactive methods for FGVC [9, 55, 27], including using perceptual similarity judgments from human users [57]. Perceptual similarity is advantageous for categorization as it does not necessitate detailed ground truth annotations such as object segmentations or part and attribute labels. Moreover, by eliminating part and attribute vocabularies, these similarity-oriented systems reduce both the burden on non-expert human users, who are not required to understand domain-specific jargon, as well as the reliance on experts, who must define these vocabularies. These factors contribute to the overall ease and flexibility in porting similarity-based categorization systems to other basic-level categories.

While similarity can be holistic in nature (*e.g.*, object utility or function, or overall shape), it can also be highly localized, for instance, when specific corresponding regions or parts of the object differ from one other. Especially at the fine-grained category level in which classes tend to be visually coherent, it is likely that the small yet important characteristics that distinguish subcategories are localizable. In these scenarios, a single metric of perceptual similarity that is observed at the object level can be overly general, and asking a user to make holistic nonlocalized similarity comparisons can be difficult.

By using localized similarity comparisons and constraining the user's view to a portion of the image, we are able to highlight certain aspects of similarity; these localized judgments tend to be easier for humans to perform than holistic similarity judgments (see Figure 1). Moreover, we can potentially reduce the effect of nuisance factors such as background noise and differing object poses. For each common region or part, we learn a separate perceptual space that captures local visual information.

In order to compare common local regions between images, we must first identify the set of relevant regions to consider, and second, we must determine spatial correspondences between regions across images and objects. For

*This work was done while the author was at UC-San Diego.

many basic-level categories, there exist field guides that specify part vocabularies for describing or discriminating categories, but these share the same weaknesses as semantic attribute vocabularies. The regions that are most useful for discrimination may not align with part semantics, and moreover, additional annotation is required to localize all the regions in the images.

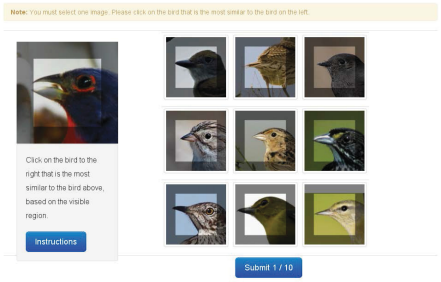
We propose using an unsupervised approach to discovering discriminative, visually coherent and roughly aligned regions [48, 13] in the dataset, which can be used to localize the similarity comparisons. This method has multiple advantages: first, we can determine spatial correspondences between images by using the discovered patches as detectors; second, the regions are by nature common in gradient appearance; and lastly, the discovered regions may provide implicit (albeit noisy) pose alignment.

Our contributions in this work are three-fold. First, we present an approach to interactive classification that leverages localized similarity comparisons and does not rely on part or attribute vocabularies. We discover a set of discriminative, localized and roughly aligned regions for this FGVC task. Second, we provide a quantitative analysis of how human users respond differently to nonlocalized versus localized perceptual similarity comparisons. Finally, we demonstrate that localized similarity comparisons are more intuitive for users to perform, and that by using independent localized metrics we can improve categorization accuracy over using a single nonlocalized metric.

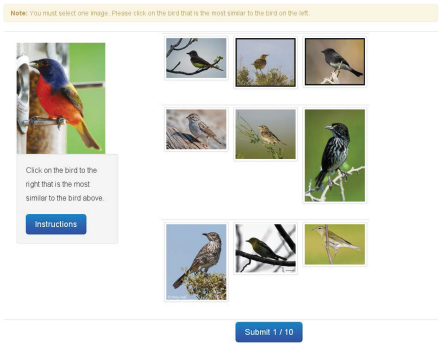
The rest of the paper is organized as follows. In Section 2, we present an overview of the relevant literature. In Section 3, we present our system. In Sections 4 and 5, we discuss implementation details and our experimental results, respectively. We conclude in Section 6.

2. Related Work

Recently, there has been increased interest in the area of fine-grained visual categorization [22, 10] with humans in the loop [9, 55, 27]. Some of this work leverages attributes that are harvested from existing resources [28, 19, 26], while others discover attributes automatically [45, 5, 15] or have humans assist in the annotation process [30, 29]. For example, humans may be asked to identify and name semantically meaningful attributes [40, 25, 34, 35], or provide justification or feedback with respect to attributes in order to improve classification accuracy [14, 43, 7, 42]. In conjunction with attribute-based methods, parts are also used to localize attributes and to assist in classification, for example by using trained one-vs.-one region-based classifiers as features [4, 3]. Often, humans are used to densely annotate these parts [8, 56] and provide pose information [20, 58]. This part vocabulary may also be generated in an automatic manner. Human-centric approaches have used partial correspondences provided by humans to extrapolate annotations to the entire dataset [35, 36]; others formulate the annotation process as a game-with-a-purpose [54, 12]. On the other hand, feature-centric approaches use image-based features to automatically discover discriminative mid-level patches [48, 13, 24] or localize attributes [15].



(a) Localized comparison



(b) Nonlocalized comparison

Figure 2. User interfaces for interactive categorization.

In this work, we focus on a part and attribute vocabulary-free approach that utilizes perceptual similarity-based comparisons rather than semantic distinctions. Within the computer vision community, several works have explored using similarity to perform classification, from the feature level [2, 26] to high-level semantic representations [59, 41, 40, 33]. We specifically use relative comparisons, which have been leveraged in categorization [46, 1, 37, 49, 47], attribute classification [41, 29], clustering [23, 6], and in improving classifiers [43, 7]. Some works capture multiple modalities of similarity rather than use a single metric [50, 38, 11]; we focus on learning an independent metric of perceptual similarity for each shared localized region across the dataset. Similar to [57], we use stochastic triplet embedding [51] to learn these metrics.

3. Approach

3.1. Problem Definition

Given a reference image x, our goal is to predict as quickly as possible the true object class c from C possible classes that fall within the same basic-level category. We do so by using both computer vision and user responses to

similarity-based questions posed by the system at test-time. Each question is shown as a display of D images, and the user is asked to make a relative judgment of similarity on the D images with respect to the image x; this perceptual measurement is recorded as response u.

Our system supports two types of similarity comparisons: nonlocalized and localized (see Figure 2). In the former, the images in the display each show the whole object. For the latter, users are asked to make a localized judgment of similarity, and all images in the grid are localized with respect to a *region* r, drawn from a set of discriminative regions $\mathcal{R}$. We define a region as a visually discriminative and recurring object part that does not have to be semantically defined or meaningful. In practice, it is a spatially localized and roughly aligned template derived from an associated descriptor (see Figure 3).

Each region for a localized comparison (or no region, in the case of a nonlocalized comparison) specifies a similarity comparison *configuration*. The set of A similarity configurations used by the system can consist of localized region-specific comparisons as well as a nonlocalized comparison.

We can represent an image x in pixel space as a vector $\mathbf{z}$ in human perceptual space. Offline, the system is provided N images annotated with subcategory labels $\{(x_i, c_i)\}_{i=1}^{N}$. We assume that we can decompose similarity into multiple similarity metrics over A different configurations. For each configuration a, we then ask human users to judge similarity with respect to it, in order to learn a separate perceptual embedding $\mathbf{Z}^a$, $a = 1 \dots A$. At test time, we observe an image x and ask a human user to draw similarity comparisons that are used to incrementally refine our probabilistic estimates of $\mathbf{z}$ and c.

Our model is based on [57], and we briefly provide an overview of this system in Section 3.2. In Section 3.3, we discuss how we extend the prior system in order to support localized similarity comparisons.

3.2. Interactive Classification System

We first discuss the interactive classification system of Wah *et al*. [57], which uses only nonlocalized similarity comparisons to learn a single perceptual metric. This corresponds to a single similarity comparison configuration and its associated embedding. For the sake of clarity, our notation below is agnostic to configuration.

Generating a perceptual metric for similarity. We represent image x as a vector $\mathbf{z}$ in this perceptual space. We first collect a set of M user similarity comparisons under this configuration; additional details on the data collection can be found in Section 4.1. A user is asked to judge the similarity of an image x with respect to a set $\mathcal{I}$ of G images, shown in a display D_t at question t (see Figure 2). From each user response u_m, where $m = 1 \dots M$, we can generate a set of triplet constraints $\mathcal{T}^m = \{(i, j, l) | x_i \text{ is more similar to } x_j \text{ than } x_l\}$, where x_i is the reference image; x_j is drawn from the set of images $\mathcal{I}_S$ that the user has selected as similar to x_i; and x_l is drawn from the set of images $\mathcal{I}_D$ that were not selected, such that $\mathcal{I}_S \cup \mathcal{I}_D = \mathcal{I}$ and $\mathcal{I}_S \wedge \mathcal{I}_D = \emptyset$.

The set $\mathcal{T}$ over N training images can be used to learn a perceptual embedding $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_N] \in \mathbb{R}^d$ for some $d \leq N$. From this embedding, we generate a similarity matrix $S \in N \times N$ with entries:

$$S_{ij} = \exp\left(-\frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{2\sigma^2}\right), \tag{1}$$

that can be used directly in our system. The scaling parameter σ is jointly learned with the user response model parameters [57].

Human-in-the-loop categorization. Recall that at test time, the goal of the system is to estimate the true class c of the reference image x as quickly as possible. As the user provides more responses, the system updates the display in an intelligent manner, and we iteratively refine probabilistic estimates of $\mathbf{z}$ and c. Let U_t be the set of user responses provided through question t. We can compute class probabilities by marginalizing over all possible locations $\mathbf{z}$:

$$\begin{aligned} p(c, U_t|x) &= \int_{\mathbf{z}} p(c, \mathbf{z}, U_t|x) d\mathbf{z} \\ &= \int_{\mathbf{z}} p(U_t|c, \mathbf{z}, x) p(c, \mathbf{z}|x) d\mathbf{z} \end{aligned} \tag{2}$$

where $p(U_t|c, \mathbf{z}, x)$ models how users respond to similarity questions, and $p(c, \mathbf{z}|x)$ is a computer vision estimate. We assume that a user's response to a similarity question is only dependent on the true location $\mathbf{z}$ of image x in perceptual space, and all answers are independent of one another.

Efficient computation. We are able to compute $p(U_t|c, \mathbf{z}, x)$ efficiently by maintaining weights $w_k^t = p(c_k, \mathbf{z}_k, U_t|x)$ for each image x_k in the training set and approximating the integral in Eq 2 as the sum of the weights of the training examples of class c:

$$p(c|x, U_t) = \frac{\sum_{k, c_k = c} w_k^t}{\sum_k w_k^t}. \tag{3}$$

Each weight w_k captures how likely $\mathbf{z}_k$ is the true location $\mathbf{z}$, and we can efficiently update w_k^{t+1} from w_k^t:

$$w_k^{t+1} \quad = \quad p(u_{t+1}|\mathbf{z}_k) w_k^t = \frac{\phi(S_{ik})}{\sum_{j \in D} \phi(S_{jk})} w_k^t, \tag{4}$$

where i is the image selected by the user at timestep $t + 1$, S_{ik} is drawn from the generated similarity matrix (Eq 1), and $\phi(\cdot)$ is a customizable function. We refer the reader to [57] for additional details on the user response model and how computer vision can be incorporated.

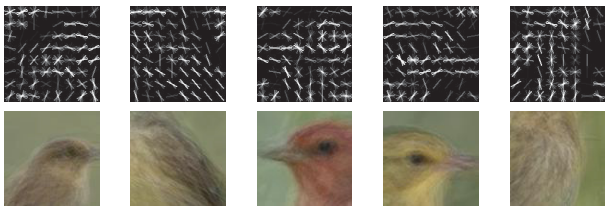


Figure 3. We discover 106 discriminative regions total (Section 3.3.1), selecting a subset of 5 representative regions to use in our experiments (Section 4.3). Each region is visualized as a HOG template (top) and the averaged image of the highest confidence positive detections for that corresponding detector (bottom).

3.3. Incorporating Localization Information

The system by Wah *et al*. [57] supports the use of multiple similarity metrics, but it does not adequately handle instance-level variations, specifically the presence or visibility of certain pose-aligned parts in the image. In this section, we describe how we automatically obtain the set of discriminative regions to localize similarity comparisons (Section 3.3.1) and how we choose which images and regions to show in the display (Section 3.3.2).

3.3.1 Discovering Discriminative Regions

In order to highlight the same localized region across images for performing localized similarity comparisons, we require instance-level region correspondences. We use the unsupervised approach of Singh *et al*. [48] to discover a set of mid-level discriminative visual representations that are localized and roughly pose aligned. At test time, we can use these templates as part detectors that are evaluated on input images in a sliding window manner. The initial candidate regions are extracted at random from uncropped images across multiple categories to ensure that regions are of sufficient resolution and of diverse scale. We found that fixing the candidate regions to lie within the object bounding box was too constraining, as the bounding boxes were often very small relative to the size of the image. The discriminative classifiers are iteratively trained on a positive set of training examples belonging to a single basic-level category, and a negative set consisting of images from all other categories, drawn from the PASCAL VOC dataset [16]. We refer the reader to Section 4.3 for additional details.

3.3.2 Question Selection and Display Model

Our augmented system must be able to: (1) select a similarity configuration at each time step to pose to the user as a question; and (2) given a particular similarity configuration, update the display accordingly. We are able to support these two additional features without adding to the overall computational complexity of the system.

Question selection. Recall that at train time, we obtain an embedding $\mathbf{Z}^a$ for each configuration $a = 1 \dots A$ using targeted similarity questions, and we generate a similarity matrix S^a from the learned embedding (Section 3.2). At test time, the similarity comparison configurations are represented as different questions, in which the system directs the user's attention to a specific region.

At each timestep t during test time, we wish to pick both a configuration a and display of images D that is likely to provide the most information gain. To do so, we identify the configuration that can produce the most balanced clustering according to the current weights w_k^t. Computation of updated class probabilities occurs identically as described in Section 3.2, with a modified update rule that replaces Eq 4:

$$w_k^{t+1} = p(u_{t+1}|\mathbf{z}_k^a) w_k^t = \frac{\phi(S_{ik}^a)}{\sum_{j \in D} \phi(S_{jk}^a)} w_k^t. \quad (5)$$

Here, we update weights w_k^{t+1} according to the similarity matrix S^a of the selected configuration a.

Updating the display. It is likely that the localized regions discovered in Section 3.3.1 may not be present in certain images; this corresponds to a low detection score for a particular region detector. As such, we modify the display model of [57] to take part presence into account.

Intuitively, for a particular region r, we wish to include images in the display that are highly likely to contain that localized region. Recall that we have a set $\mathcal{R}$ of discriminative regions. For a given image x_k in the training set, we model the probability the region $r \in \mathcal{R}$ is present in x_k as $p(v_k|r, x_k)$. In practice, this is determined by applying a sigmoid function to the output of the region detector. The γ parameter is learned on a validation set [44].

In selecting images for the display, we employ the approximate solution described in [57, 18, 21], which groups the images into clusters to ensure that each image in the display is equally likely to be selected, maximizing the information gain in terms of the entropy of $p(c, \mathbf{z}_k, U_t|x)$. For the display, we thus pick the image within the cluster with the highest mass as weighted by the region presence probability $w_k^t p(v_k|r, x_k)$.

4. Implementation Details

4.1. Dataset and Data Collection

We perform our experiments on the fine-grained CUB-200-2011 dataset [56], which consists of 200 subcategories of bird species with roughly 60 images per class. We assume that we are provided with ground truth object bounding boxes in both training and testing, and we maintain the specified train/test split.

For learning the perceptual metrics, we collect similarity comparisons using the crowdsourcing workplace Amazon

Mechanical Turk. Collecting similarity judgments densely over all training images for each region would be an expensive and costly process; instead, we sample images for the displays from the distribution $p(v_k|r, x_k)$, such that noisy detections with low $p(v_k|r, x_k)$ are less likely to be selected for annotation. For each region r, we collect localized similarity comparisons using the GUI in Figure 2(a), which uses a 3×3 grid display of $G = 9$ images. Some context around each region is shown.

4.2. Computer Vision Features and Learning

In order to compare to previous work, we initialize our computer vision estimate of class probabilities using the same setup as [57], with multiclass 1-vs-all SVMs [17] trained on color/grayscale SIFT features and color histograms extracted from the uncropped images. We also compare to a method that uses Fisher vector encodings (FVs) with features extracted from the object bounding boxes, which has been demonstrated to improve FGVC accuracy over other computer vision algorithms [22]. We extract SIFT descriptors and use a 256-visual words GMM, applying an $L2$-normalization on the Fisher vectors and learning a linear SVM with VLFEAT [53], yielding 34.76% average classification accuracy on the test set, compared to 18.9% with SIFT/color features.

4.3. Discriminative Region Vocabulary

To generate the discriminative regions set, we only keep discovered patches that have sufficient overlap (50%) with the ground truth object bounding box [31]. This eliminates many noisy detections that fire in the image background, resulting in 106 localized and roughly aligned regions.

We also wish to ensure sufficient diversity in the regions used; consequently, we apply agglomerative clustering to reduce the set of 106 discovered regions to 23 region clusters. These region clusters are fairly noisy, and semantically aligned regions are not always grouped together due to translation and scale variance. We thus manually select 5 diverse and representative regions from different clusters to comprise $\mathcal{R}$ and to use in our experiments (see Figure 3). In practice, these regions can be selected in a more automated manner, for example using an MTurk pipeline.

5. Experiments

In Section 5.1, we describe how the embeddings are generated. In Section 5.2, we compare human perception differences between localized and nonlocalized similarity judgments. In Section 5.3, we present our interactive classification results.

5.1. Embedding Generation

For each localized region (Figure 3), we generate triplets from similarity comparisons (Section 4.1) in order to learn an independent localized embedding of N nodes and of dimensionality d for each region r. The comparisons are collected at the instance level, and for each embedding, we pool over instances in each class, such that we obtain an embedding of $N = C = 200$ and $d = 10$ [57]. This enables us to generate a similarity matrix $S^r \in C \times C$ for each region r. The metric is learned independently from all other regions; see Figure 4 for visualizations of the embeddings.

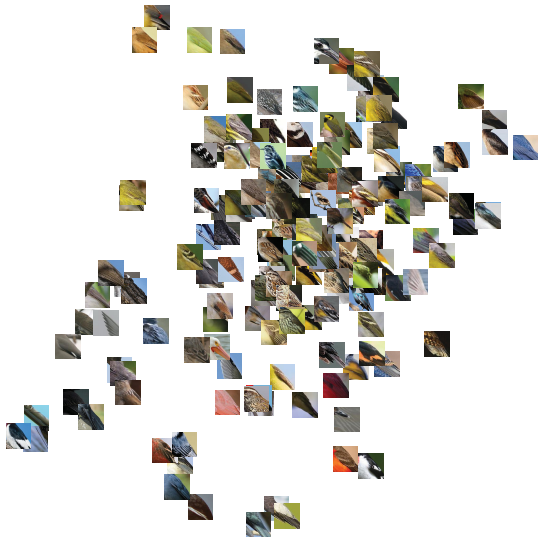


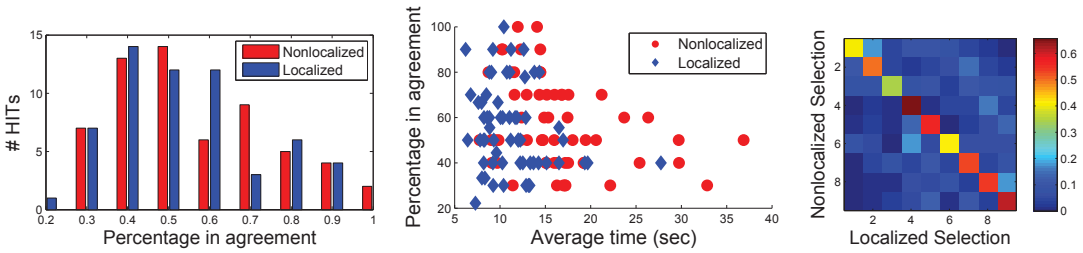
Figure 4. We visualize the first two dimensions of the embedding for one localized region (R13) used in our experiments.

This pooling step mitigates the effects of noise in both user similarity responses and region detection, and we find that we do not need to filter any noisy user responses from training in order to learn the embeddings. By pooling over classes, we assume that the visual appearance of parts are coherent within a subcategory; in reality, however, there is intraclass variation due to differences in gender, age, season, etc. While we do not directly address this, our user response model is able to account for noise in user responses.

5.2. Human Perception of Localized Similarity

We first observe empirically how users respond differently to localized compared to nonlocalized similarity questions. We generate 20 unique questions, each of which consists of 10 images total: a reference image and a grid of $G = 9$ images. Each question is seen by up to 10 AMT workers. The 200 images in the questions are selected from the top-scoring detections across the dataset for Region 1, Region 21, and Region 39 (see Figure 3).

We create two experiments from the set of 20 questions. One consists of localized questions only, in which the detected region is highlighted in the image (Figure 2(a)). The second experiment consists of the same set of questions, but the images are shown to the user as the full uncropped version (Figure 2(b)). Across both experiments, the images in



(a) (b) (c)

Figure 5. *Comparing human perception of nonlocalized vs. localized similarity.* 5(a): Histogram of HITs with a certain maximum percentage agreement (MPA) among 10 AMT worker responses; 5(b): MPA vs. average worker time per question; 5(c): the co-occurrence rates of user-selected image locations in the 3×3 grid, enumerated $1-9$.

the grid appear in the same position, and the user is asked to select a single image in the grid that is most similar to the reference image. The only variable that changes between experiments is how the images are displayed to the user. Due to how the discriminative regions are discovered, both experiments show images that are roughly pose aligned. We present our results in Figure 5.

Localized similarity comparisons require less human effort. We observe in Figure 5(b) the relationship between user consistency and response time. Each point corresponds to a single question. We plot the maximum percentage agreement (MPA) of the 10 users who agree on a single image in the grid, versus the average response time over those users. On average, it takes a human user less time (and with lower variance) to answer a localized comparison (11.35 ± 10.17 sec), compared to 16.36 ± 14.31 sec for a nonlocalized comparison.

Users answer both localized and nonlocalized questions with similar consistency. In Figure 5(a), we plot the distribution of tasks according to the MPA. Both localized and nonlocalized similarity questions have comparable average MPA across questions (0.54 vs. 0.56, respectively), suggesting that users answer the two types of questions with similar levels of consistency. With localized regions, there still exist multiple dimensions upon which to judge similarity, such as color, shape, and pattern. As such, localization does not necessarily remove ambiguities, but does make the comparison task easier to perform.

Responses to localized questions yield different information about similarity. We present in Figure 5(c) the co-occurrence rates of selected image locations in the 3×3 grid (1-9, enumerated in left-to-right, top-to-bottom order) for corresponding nonlocalized and localized questions. Selections are normalized by row. Users select the same image as the most similar for both nonlocalized and localized questions only 50.73% of the time on average, indicating that a localized similarity response provides different visual similarity information to the system. We do note that worker noise and bias can affect their responses [32]; for example, workers tend to click on the lower-left portion of the grid, as it is closer to the button to advance to the next question.

5.3. Interactive Categorization

We show our results on interactive classification in Figure 6; qualitative examples are presented in Figure 7. At test time, we use an interface similar to that used in training (Figure 2(a)), with the primary difference being how the reference image is displayed. The region detections in uncropped test images can be noisy, and we wish to avoid highlighting an erroneous detection to the user. Instead, we show the nonlocalized reference image, and we assume that, with some cost in human effort, the user is able to mentally localize and align the corresponding region, based on the localized region highlighted in the grid images.

Similar to [57], we use simulated user responses that allow us to compare to previous work more readily as well as explore different parameter choices. We use a model for user behavior that accounts for noisy responses, estimating parameters on a validation set of real human responses. We refer the reader to [57] for details on the user model. Our experimental setup and performance metrics are the same as [55, 57], in which the user can verify perfectly the highest probability class, and we evaluate our system based on the average number of questions a user must answer per test image to classify it correctly.

It is advantageous to use localized and nonlocalized metrics together. In Figure 6(a), we compare performance with simulated noisy users to the system presented in [57], which uses a single nonlocalized class similarity metric (*Wah2014*), as well as to previous baselines from [57]: an interactive classification system that uses part-localized computer vision algorithms and poses semantic part click and binary attribute questions (*Wah2011*) [55]; an implementation of a relevance feedback system that uses a feature-based $L1$-distance metric (*Ferecatu2009*) [21]; and a baseline derived from classification scores alone, in which the user moves down the ranked list of classes to verify the correct class (*Ranked by CV*). Class probabilities are

initialized using the computer vision algorithms based on SIFT/color histograms.

By combining localized and nonlocalized metrics, we are able to classify the test images with 9.85 questions on average, compared to 9.99 by using localized metrics only and 11.53 from using the nonlocalized metric. In Figure 6(b), we observe a similar trend when we use FV-based computer vision estimates for initializing per-class probabilities; using both types of metrics results in 0.44 less questions on average than using only localized metrics. We emphasize that this performance gain is further exaggerated when we consider the observation that localized comparisons take on average 5.01 sec less time to perform than nonlocalized comparisons (Section 5.2).

We also compare to the *Ranked by CV* baseline using the Fisher vector encoding. This baseline outperforms our system initially but fails on more difficult images, whereas our similarity-based approach is able to ultimately identify the correct class.

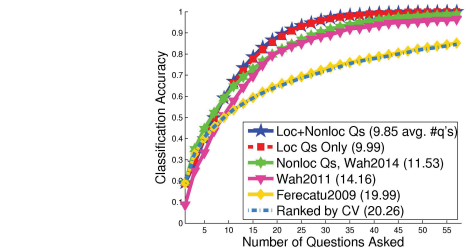
Localized comparisons are more informative than nonlocalized comparisons. In general, our interactive categorization system will tend to ask users to make localized comparisons in the beginning, as these questions provide the most expected information gain. As the per-class probability estimates are refined, the system will ask more nonlocalized similarity questions.

Some localized regions are more useful for categorization than others. In Figure 6(c) we present categorization results using the localized metrics separately. We note that using the localized metric for Region 13 outperforms using the other localized metrics. This may suggest that the visual representation captured by Region 13, visualized in Figure 4, is particularly useful for discriminating bird species. Nevertheless, it is still be beneficial to use a combination of regions, as not all regions will be present in all the images. For example, using Region 13 alone produces a boost in average categorization accuracy initially for the first 15 questions; after that point, other localized metrics become more informative.

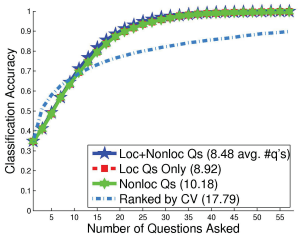
6. Conclusion

We present in this work a part and attribute vocabulary-free approach for interactive categorization that uses localized perceptual similarity metrics. These metrics are learned independently, even if the regions are overlapping on the object. In the future, we can take spatial dependencies into account in learning these localized metrics.

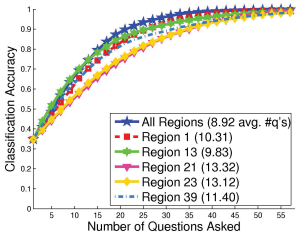
Performance of the system is affected significantly by noisy detections, which impact how accurately a user can judge localized similarity. To alleviate this and improve classification accuracy, we can consider leveraging human feedback to clean up poor detections or to aid in the selection of discriminative regions to incorporate in the system.



(a) SIFT/color histograms



(b) Fisher vectors (FVs)



(c) Local metrics (FVs)

Figure 6. *Interactive categorization.* 6(a): Using both localized and nonlocalized metrics outperforms using either type of metric alone. We compare to prior baselines from [57]. 6(b): We observe performance when the initial class probability estimates are improved by using Fisher vectors. 6(c): We compare performance using each localized metric separately.

Because our system is able to discover and detect regions in an unsupervised manner, it requires minimal manual intervention. By having humans perform the aforementioned steps as an automated crowdsourced process, we will have all the necessary components to deploy this system as a standalone interactive categorization pipeline. The only required input to this system would be a dataset of images with class labels, enabling this system to potentially be extended to new domains easily and automatically.

7. Acknowledgments

The authors thank the reviewers for their helpful feedback. This work is supported by the Google Focused Research Award and the National Science Foundation Graduate Research Fellowship under Grant No. DGE0707423.

References

[1] S. Agarwal, J. Lim, L. Zelnik-Manor, P. Perona, D. J. Kriegman, and S. Belongie. Beyond pairwise clustering. In *CVPR*, 2005.

[2] B. Babenko, S. Branson, and S. Belongie. Similarity metrics for categorization. In *CVPR*, 2009.

[3] T. Berg and P. N. Belhumeur. How do you tell a blackbird from a crow? In *ICCV*, 2013.

[4] T. Berg and P. N. Belhumeur. POOF : Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013.

[5] T. Berg, A. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.

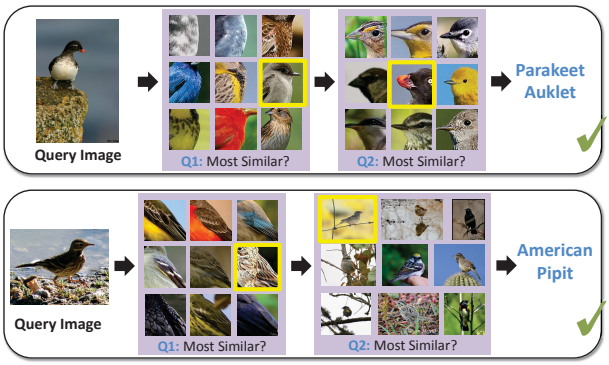


Figure 7. Qualitative examples using only the 5 localized similarity metrics (top) and using the localized metrics along with a nonlocalized metric (bottom).

[6] A. Biswas and D. Jacobs. Active image clustering: Seeking constraints from humans to complement algorithms. In *CVPR*, 2013.

[7] A. Biswas and D. Parikh. Simultaneous active learning of classifiers & attributes via relative feedback. In *CVPR*, 2013.

[8] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[9] S. Branson et al. Visual recognition with humans in the loop. In *ECCV*, 2010.

[10] Y. Chai, V. Lempitsky, and A. Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *ICCV*, 2013.

[11] S. Changpinyo, K. Liu, and F. Sha. Similarity component analysis. In *NIPS*, 2013.

[12] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, 2013.

[13] C. Doersch et al. What makes Paris look like Paris? *ACM Transactions on Graphics*, 31(4):1–9, July 2012.

[14] J. Donahue and K. Grauman. Annotator rationales for visual recognition. In *ICCV*, 2011.

[15] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering Localized Attributes for FGVC. In *CVPR*, 2012.

[16] M. Everingham et al. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[17] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR. *JMLR*, 2008.

[18] Y. Fang and D. Geman. Experiments in mental face retrieval. In *AVBPA*, 2005.

[19] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[20] R. Farrell, O. Oza, V. I. Morariu, T. Darrell, and L. S. Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *ICCV*, 2011.

[21] M. Ferecatu and D. Geman. A statistical framework for category search from a mental picture. *TPAMI*, 2009.

[22] E. Gavves, B. Fernando, C. Snoek, A. Smeulders, and T. Tuytelaars. Fine-grained categorization by alignments. In *ICCV*, 2013.

[23] R. Gomes, P. Welinder, A. Krause, and P. Perona. Crowdclustering. In *NIPS*, 2011.

[24] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013.

[25] A. Kovashka, S. Vijayanarasimhan, and K. Grauman. Actively selecting annotations among objects and attributes. In *ICCV*, 2011.

[26] N. Kumar et al. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[27] N. Kumar et al. Leafsnap: A computer vision system for automatic plant species identification. In *ECCV*, 2012.

[28] C. Lampert et al. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[29] E. Law et al. Human computation for attribute and attribute value acquisition. In *CVPR FGVC Workshop*, 2011.

[30] E. Law and L. von Ahn. Input-agreement: A new mechanism for data collection using human computation games. In *CHI*, 2009.

[31] Y. J. Lee et al. Style-aware mid-level representation for discovering visual connections in space and time. In *CVPR*, 2013.

[32] G. Little. Top, middle, or bottom? `http://groups.csail.mit.edu/uid/deneme/?p=27`, 2009.

[33] S. Ma, S. Sclaroff, and N. Ikizler-Cinbis. Unsupervised Learning of Discriminative Relative Visual Attributes. In *ECCV Workshop on Parts and Attributes*, 2012.

[34] S. Maji. Discovering a lexicon of parts and attributes. In *ECCV Workshop on Parts and Attributes*, 2012.

[35] S. Maji and G. Shakhnarovich. Part annotations via pairwise correspondence. In *Human Computation Workshop*, 2012.

[36] S. Maji and G. Shakhnarovich. Part discovery from partial correspondences. In *CVPR*, 2013.

[37] B. McFee and G. Lanckriet. Metric learning to rank. In *ICML*, June 2010.

[38] B. McFee and G. Lanckriet. Learning multi-modal similarity. *JMLR*, 2011.

[39] M. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICCVGIP*, 2008.

[40] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011.

[41] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[42] D. Parikh and K. Grauman. Implied feedback: Learning nuances of user behavior in image search. In *ICCV*, 2013.

[43] A. Parkash and D. Parikh. Attributes for classifier feedback. In *ECCV*, 2012.

[44] J. Platt. Probabilities for SV machines. In *NIPS*, 1999.

[45] M. Rohrbach et al. What helps where – and why? semantic relatedness for knowledge transfer. In *CVPR*, 2010.

[46] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In *NIPS*, 2003.

[47] A. Shrivastava, S. Singh, and A. Gupta. Constrained semi-supervised learning using attributes and comparative attributes. In *ECCV*, 2012.

[48] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

[49] O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. Kalai. Adaptively learning the crowd kernel. In *ICML*, 2011.

[50] L. van der Maaten and G. Hinton. Visualizing non-metric similarities in multiple maps. *ML*, 2012.

[51] L. van der Maaten and K. Weinberger. Stochastic triplet embedding. In *MLSP*, 2012.

[52] A. Vedaldi et al. Understanding objects in detail with fine-grained attributes. In *CVPR*, 2014.

[53] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library. `http://www.vlfeat.org/`, 2008.

[54] L. von Ahn, R. Liu, and M. Blum. Peekaboom: A game for locating objects in images. In *CHI*, pages 55—64, 2009.

[55] C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass rec. and localization with humans in the loop. In *ICCV*, 2011.

[56] C. Wah et al. Caltech-UCSD Birds-200-2011. Technical Report CNS-TR-2011-001, Caltech, 2011.

[57] C. Wah et al. Similarity comparisons for interactive fine-grained categorization. In *CVPR*, 2014.

[58] N. Zhang et al. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV*, 2013.

[59] G.-T. Zhou, T. Lan, W. Yang, and G. Mori. Learning class-to-image distance with object matchings. In *CVPR*, 2013.